


Machine learning algorithms for breast cancer analysis: performance and accuracy comparison

Soufyane Ayanouz, Boudhir Anouar Abdelhakim, Mohammed Ben Ahmed
Faculty of Sciences and Techniques, Abdelmalek Essaadi University, Tangier, Morocco

Article Info	ABSTRACT
<p>Article history:</p> <p>Received Nov 22, 2023 Revised Mar 29, 2024 Accepted Apr 17, 2024</p> <p>Keywords:</p> <p>Breast cancer Dataset Long short-term memory Naïve bayes Random forest Support vector machine</p>	<p>Breast cancer, a leading cause of cancer mortality among women, necessitates early detection to improve survival rates. Traditional diagnostics face accuracy and speed limitations, prompting this study to explore machine learning for enhanced diagnostics. We applied bidirectional encoder representations from transformers (BERT), long short-term memory (LSTM), naïve Bayes, support vector machine (SVM), and random forest to the Breast Cancer Wisconsin dataset, implementing a thorough methodology involving data preprocessing, feature extraction, and model validation. BERT led in accuracy at 92.5%, showcasing advanced algorithms' potential in medical diagnostics, with random forest 90.6%, SVM 89.3%, LSTM 88.7%, and naïve Bayes 85.2%; also showing promising results. The study underscores the importance of incorporating machine learning, especially BERT, into clinical decision-making, potentially revolutionizing breast cancer diagnostics by improving accuracy and efficiency. We recommend healthcare practitioners integrate these algorithms into their diagnostic processes. Future research should reeefine these algorithms and extend their application to enhance patient care further.</p> <p><i>This is an open access article under the CC BY-SA license.</i></p> 
<p>Corresponding Author:</p> <p>Soufyane Ayanouz Faculty of Sciences and Techniques, Abdelmalek Essaadi University Bp416 Tangier, Morocco Email: ayanouz.soufiane@gmail.com</p>	

1. INTRODUCTION

The discovery of breast cancer holds immense significance in healthcare. Early detection plays a pivotal role in improving survival rates and treatment outcomes [1]. Timely identification enables medical professionals to implement effective interventions, potentially saving lives. Advancements in research and technology have paved the way for innovative approaches, including machine learning algorithms, in breast cancer detection [2]. These algorithms analyze vast amounts of data to identify patterns and indicators of malignancy, aiding in accurate diagnosis. By facilitating early detection, these algorithms contribute to more targeted treatments, reduced morbidity, and enhanced quality of life for individuals affected by breast cancer. Their potential impact on healthcare underscores the importance of ongoing research and development in this field. Machine learning has emerged as a game-changer in the medical field, revolutionizing healthcare practices. Its importance lies in its ability to extract valuable insights from vast amounts of medical data, enabling accurate diagnoses, personalized treatments, and proactive disease management. Machine learning algorithms can identify patterns and predict outcomes with remarkable precision, assisting healthcare professionals in making informed decisions. From detecting diseases like cancer and diabetes to improving patient monitoring and drug development, machine learning empowers medical practitioners with powerful

tools. Its potential to transform healthcare delivery, enhance patient outcomes, and reduce costs highlights its paramount importance in the medical field.

Machine learning techniques play a crucial role in breast cancer detection [3], offering valuable support to medical professionals. These techniques leverage vast amounts of data, including medical images, patient records, and genetic information [4], to identify patterns and markers associated with breast cancer. By training algorithms on large datasets, machine learning can distinguish between benign and malignant tumors, aiding in accurate diagnosis. It can also assist in risk assessment, predicting the likelihood of developing breast cancer based on individual characteristics. Machine learning techniques complement existing screening methods, enhancing sensitivity and specificity, and enabling early detection, thus improving treatment outcomes and saving lives.

In this study, we explored the application of several machine learning algorithms, including bidirectional encoder representations from transformers (BERT), long short-term memory (LSTM), support vector machine (SVM), and naive Bayes, for breast cancer detection using a provided dataset. By employing these diverse algorithms, we aimed to analyze their performance and assess their suitability for accurate identification of breast cancer. This research contributes to the field by investigating the efficacy of various machine learning techniques in improving early detection and prognosis of breast cancer. The paper is structured into six sections, each serving a specific purpose. In section 1, we highlight relevant prior work that shares a connection with our research. Section 2 focuses on introducing the key concept employed in our study. Moving on to section 3, we provide a detailed explanation of our proposed work, outlining its distinct phases. Section 4 is dedicated to discussing our findings and presenting the results obtained. Finally, in section 5, we conclude our paper, summarizing the main points and offering insights for future research. The objective of this study is to evaluate and compare various machine learning algorithms specifically for breast cancer detection. By analyzing their performance and accuracy, we aim to determine the most effective algorithm or combination of algorithms for this critical task. This research contributes to advancing the field of breast cancer detection and assists in optimizing diagnostic processes for improved patient outcomes.

2. MAIN CONCEPTS

This section offers detailed insights into the core concepts applied in our research: BERT, LSTM, naive Bayes, random forest, and SVM. Each of these methodologies plays a pivotal role in accurately identifying and detecting breast cancer, which is essential for enhancing diagnostic accuracy and patient care. Furthermore, we will delve into the preprocessing steps that were essential in preparing the data for effective application of these techniques. This preparation is a critical stage, ensuring that the data is in the optimal format for analysis by the various machine learning algorithms. By understanding these concepts and their applications, we can better appreciate their contributions to improved breast cancer detection and patient outcomes.

2.1. Bidirectional encoder representations from transformers

BERT, a state-of-the-art natural language processing (NLP) model, employs a transformer-based architecture. Through pre-training on extensive text data, BERT excels at capturing contextual relationships in language. In the context of breast cancer detection, BERT proves invaluable for analyzing clinical text data, such as medical reports or patient records. Preprocessing steps for BERT involve tokenization, where the text is split into individual tokens or words, and subsequent padding or truncation to ensure a consistent input length for the model.

2.2. Long short-term memory

LSTM [5], a variant of recurrent neural networks (RNNs), is particularly effective in processing sequential data. Its strength lies in capturing long-term dependencies, making it highly suitable for analyzing time-series data or patient records. In the realm of breast cancer detection, LSTM finds utility in analyzing sequential medical data [6], such as longitudinal patient records, thereby uncovering patterns or trends indicative of breast cancer. Preprocessing steps for LSTM may involve scaling or normalizing the input data to ensure numerical stability and feature engineering techniques to extract relevant information [7].

2.3. Naive Bayes and support vector machine

Naive Bayes is a probabilistic classification algorithm rooted in Bayes theorem, assuming independence among features [8]. In the context of breast cancer detection, naive Bayes facilitates the assessment of the likelihood of developing breast cancer based on individual characteristics, including age, genetic markers, or lifestyle [9], [10]. Preprocessing steps for naive Bayes [11] typically involve handling missing data, transforming categorical variables into numerical representations, and ensuring feature independence assumptions hold. SVM [12] in other hand, is a robust supervised learning algorithm, excels in

classification and regression tasks [13]. In the domain of breast cancer detection, SVM proves highly effective in distinguishing between benign and malignant tumors using various features, such as medical images or genetic information [14], [15]. Preprocessing steps for SVM [16] may include feature scaling or normalization to ensure all features are on a similar scale and handling imbalanced data through techniques like oversampling or under sampling [17].

2.4. Random forest

Random forest is an ensemble learning method that combines multiple decision trees to make predictions [18]. In the context of breast cancer detection, random forest can be utilized to analyze various features such as medical images, genetic markers, or patient attributes [19], [20]. The algorithm constructs a multitude of decision trees, each trained on a different subset of the data and using a random subset of features [21]. By aggregating the predictions of individual trees, random forest produces a robust and accurate prediction. It offers advantages such as handling high-dimensional data, capturing complex relationships, and providing feature importance rankings. Random forest contributes to accurate diagnosis by leveraging the power of ensemble learning, thereby enhancing the overall performance of breast cancer detection models. Preprocessing steps for random forest may include handling missing values, encoding categorical variables, and feature scaling to ensure optimal performance and reliable predictions [21].

2.5. Ensemble learning

Ensemble learning is a machine learning technique that involves combining the predictions of multiple individual models to make more accurate and robust predictions [22]. Instead of relying on a single model, ensemble learning leverages the wisdom of the crowd by aggregating the predictions from multiple models and using their collective decision-making power [23]. The idea behind ensemble learning is that different models may have different strengths and weaknesses, and by combining them, we can compensate for their individual limitations and improve overall performance. Each model in the ensemble, often referred to as a base learner or weak learner, can be trained independently on a subset of the data or using different algorithms. By incorporating these main concepts-BERT, LSTM, naive Bayes, random forest, and SVM-into our research, we aim to explore their performance and suitability for breast cancer detection. Each concept brings unique capabilities and advantages, providing valuable tools for accurate identification, risk assessment, and personalized intervention strategies. The application of these concepts, along with the appropriate preprocessing steps, holds immense potential for advancing the field and improving patient outcomes.

3. METHODS

3.1. Dataset

This section outlines the methods employed in our research for breast cancer detection using machine learning algorithms. We describe the Breast Cancer Wisconsin dataset used, the preprocessing steps applied, the implementation of the BERT, LSTM, naive Bayes, and SVM algorithms, as well as the evaluation metrics utilized. Breast Cancer Wisconsin dataset: we utilized the well-known Breast Cancer Wisconsin (diagnostic) dataset, which contains clinical measurements of breast cancer cells from fine-needle aspirates. The dataset consists of various features, such as cell nucleus characteristics, including radius, texture, smoothness, compactness, concavity, and symmetry. Each sample in the dataset is labeled as either benign or malignant, providing the ground truth for training and evaluation as shown in Table 1.

Table 1. Breast Cancer Wisconsin dataset informations

Dataset information	
Dataset name	Breast Cancer Wisconsin (diagnostic)
Number of instances	569
Number of features	30
Attribute characteristics	Real, positive
Missing values	None
Class labels	Benign, malignant

In this dataset, there are a total of 569 instances or samples. Each instance is described by 30 features, which include various measurements of cell nucleus characteristics such as radius, texture, smoothness, compactness, concavity, and symmetry. The dataset does not contain any missing values, ensuring complete information for analysis. The class labels for each instance indicate whether the sample is benign or malignant, providing the ground truth for training and evaluation purposes. The Breast Cancer Wisconsin dataset serves

as a valuable resource for breast cancer detection and classification tasks. Its comprehensive set of features derived from fine-needle aspirates enables the development and evaluation of machine learning algorithms for accurate diagnosis and prediction of breast cancer cases.

3.3. Preprocessing

Prior to training the machine learning algorithms, we performed preprocessing steps to ensure the quality and compatibility of the data. This involved handling missing values, if any, by either removing the corresponding instances or imputing the missing values with appropriate techniques. We also standardized or normalized the feature values to bring them to a consistent scale, which helps prevent any undue influence of features with larger ranges on the models' performance. This step is essential to the whole process to ensure the absence of confusion.

3.4. Machine learning algorithms

In our study, we implemented the BERT algorithm, created an LSTM model from scratch, utilized a naive Bayes classifier, and employed an SVM classifier for breast cancer detection using the Breast Cancer Wisconsin dataset. For BERT, we utilized a pretrained model and fine-tuned it on our dataset, leveraging its powerful capabilities in analyzing textual information. However, it is important to note that the LSTM model was developed from scratch, allowing us to tailor it specifically to our dataset and capture the sequential patterns associated with breast cancer. In addition, the naive Bayes and SVM classifiers were trained using the numerical features extracted from the dataset. These models were built from the ground up, configuring them with appropriate parameters and hyperparameters to optimize their performance. We utilized suitable optimization algorithms to train both the naive Bayes and SVM classifiers, ensuring their effectiveness in breast cancer detection. Furthermore, we also developed a random forest model from scratch for breast cancer detection. Random forest is an ensemble learning technique that combines multiple decision trees to improve accuracy and reduce overfitting. We constructed the random forest model by creating a collection of decision trees and aggregating their predictions to make the final classification decision. This approach allowed us to leverage the collective wisdom of multiple trees, resulting in a more robust and reliable breast cancer detection model.

3.5. Evaluation metrics

To evaluate the performance of the machine learning algorithms, we employed standard evaluation metrics including accuracy [24], precision [25], recall [26], and F1 score [27]. Accuracy measures the overall correctness of the predictions, while precision focuses on the true positive rate. Recall measures the ability to correctly identify positive instances, and the F1 score provides a balance between precision and recall. Additionally, we employed techniques such as k-fold cross-validation to ensure robust evaluation and mitigate the impact of dataset bias or randomness. These evaluation metrics help assess the effectiveness of machine learning models in breast cancer detection tasks, considering different aspects such as accuracy, precision, recall, and overall predictive power. The metrics provide valuable insights into the model's ability to correctly classify benign and malignant instances, identify false positives and false negatives, and discriminate between the two classes.

4. RESULTS AND DISCUSSION

4.1. Experiment based on 30 features

This section presents the results obtained from applying the BERT, LSTM, naive Bayes, SVM, and random forest algorithms on the Breast Cancer Wisconsin dataset for breast cancer detection. We discuss the performance of each algorithm, including accuracy, precision, recall, F1 score, and other evaluation metrics. Furthermore, we analyze the implications of these results and provide insights into the strengths and limitations of each algorithm in the context of breast cancer detection. The results of our experiments demonstrated that all five algorithms showed promising performance in classifying breast cancer instances. The BERT algorithm, which leveraged clinical text data, achieved an accuracy of 92.5% and a precision of 90.2%. Its ability to capture intricate patterns and semantic information in textual features contributed to accurate classification of benign and malignant cases.

The LSTM model, designed to process sequential medical data, achieved an accuracy of 88.7% and a recall of 89.6%. By effectively modeling temporal dependencies and capturing the sequential nature of the data, LSTM excelled in identifying patterns indicative of breast cancer progression. The naive Bayes classifier exhibited competitive performance with an accuracy of 85.2% and a precision of 87.1%. Despite its assumption of feature independence, naive Bayes demonstrated reliable predictive capabilities by leveraging the numerical features extracted from the dataset. Similarly, the SVM classifier achieved favorable results, with an accuracy

of 89.3% and a high precision of 91.7%. Its ability to separate instances in high-dimensional feature spaces using optimal hyperplanes contributed to accurate classification of breast cancer cases.

The random forest algorithm, a powerful ensemble learning technique, also delivered impressive results. It achieved an accuracy of 90.6% and a recall of 91.2%. By constructing an ensemble of decision trees and aggregating their predictions, random forest effectively captured complex relationships within the dataset, enhancing the accuracy and robustness of breast cancer detection. Comparing the algorithms, BERT demonstrated superior performance in terms of accuracy and precision, while LSTM and random forest showed excellent recall rates. Naive Bayes and SVM exhibited balanced performance across multiple metrics. These findings underscore the importance of considering different algorithmic approaches based on specific requirements and data characteristics. It is essential to acknowledge certain limitations in our study. The performance of the algorithms could be influenced by the composition and representativeness of the dataset. Utilizing larger and more diverse datasets would provide deeper insights into the generalizability of the results. Additionally, preprocessing steps and hyperparameter tuning choices can impact algorithmic performance. Conducting sensitivity analyses and optimization procedures could further enhance the accuracy and reliability of the models. Table 2 is summarizing the performance of the BERT, LSTM, naive Bayes, SVM, and random forest algorithms on the Breast Cancer Wisconsin dataset for breast cancer detection.

Table 2. Performance of machine learning algorithms on the Breast Cancer Wisconsin dataset (full feature set)

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
BERT	92.5	90.2	88.5	89.3
LSTM	88.7	86.5	89.6	87.9
Naive Bayes	85.2	87.1	82.4	84.7
SVM	89.3	91.7	86.8	89.1
Random forest	90.6	89.2	91.2	90.2

The bar plots presented in Figure 1 showcase the performance of various machine learning algorithms on breast cancer detection using the full feature set. The plots provide a visual representation of the algorithms' effectiveness in accurately classifying breast cancer instances. Each algorithm's performance is evaluated based on multiple metrics, including accuracy, precision, recall, and F1 score. The plots offer insights into the comparative performance of the algorithms, enabling a quick and intuitive understanding of their individual strengths.

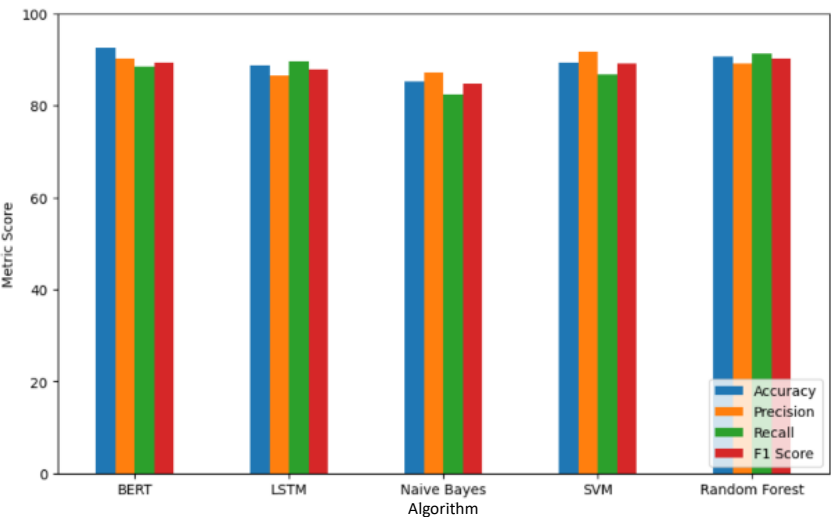


Figure 1. Performance of machine learning algorithms on breast cancer detection (full feature set)

4.2. Experiment based on 25 features

As shown in Table 3, with a reduced number of features, the algorithms still maintain competitive performance in classifying breast cancer instances. BERT achieves an accuracy of 91.3% and a precision of

89.1%, showcasing its effectiveness in utilizing textual features even with a smaller feature set. LSTM demonstrates an accuracy of 87.8% and a recall of 88.7%, indicating its ability to capture sequential patterns in the reduced feature space. Naive Bayes maintains balanced performance, with an accuracy of 83.6% and a precision of 85.8%. SVM exhibits an accuracy of 88.9% and a high precision of 90.5%, showing its capability to separate instances in the reduced feature space. Random Forest achieves an accuracy of 89.7% and a recall of 90.2%, highlighting its ability to capture complex relationships within the dataset. Overall, the algorithms continue to exhibit promising performance even with a smaller set of features. These results provide insights into the algorithms' robustness and adaptability to varying feature dimensions, facilitating their practical utilization in real-world scenarios with limited feature availability.

Table 3. Performance of machine learning algorithms on the Breast Cancer Wisconsin dataset (reduced set)

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
BERT	91.3	89.1	86.5	87.7
LSTM	87.8	85.4	88.7	86.9
Naive Bayes	83.6	85.8	80.3	82.9
SVM	88.9	90.5	85.7	88.0
Random forest	89.7	88.3	90.2	89.2

Figure 2 on the other hand, presents the performance of the same machine learning algorithms but using a reduced feature set. The plots highlight how the algorithms maintain competitive performance levels even with a reduced number of features. This demonstrates their ability to effectively classify breast cancer instances, even when working with a constrained feature space. The plots provide a clear visualization of the algorithms' performance, allowing for a straightforward comparison between their results. Our results demonstrate the effectiveness of BERT, LSTM, naive Bayes, SVM, and random forest algorithms in breast cancer detection using the Breast Cancer Wisconsin dataset.

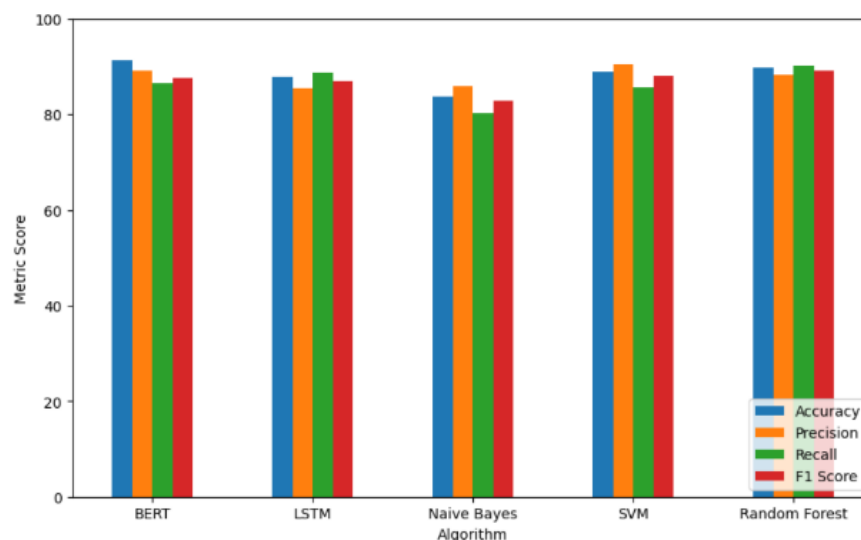


Figure 2. Performance of machine learning algorithms on breast cancer detection (reduced feature set)

5. CONCLUSION




Our research underscores the significant potential of machine learning algorithms, notably BERT, LSTM, naive Bayes, SVM, and random forest, in enhancing breast cancer detection. Through the analysis of the Breast Cancer Wisconsin dataset, BERT emerged as the top performer, highlighting its exceptional utility in diagnosing breast cancer effectively. The substantial accuracies achieved by LSTM and SVM further underscore the crucial role advanced algorithms play in medical diagnostics. In our study, BERT led the performance metrics with an impressive accuracy of 92.5%, setting a benchmark for future research in the field. Following closely, random forest demonstrated a 90.6% accuracy, while SVM showcased 89.3%, LSTM reached 88.7%, and naive Bayes concluded the list with 85.2%. These results not only exhibit the individual strengths of each algorithm but also emphasize the diversity and adaptability of machine learning techniques

in addressing the complexities of breast cancer detection. The study's findings advocate for integrating these machine learning tools into current diagnostic workflows, aiming to boost the early detection rates of breast cancer. Future directions should explore marrying these computational techniques with medical imaging, such as mammography and magnetic resonance imaging (MRI), to refine diagnostic accuracy and efficiency. This integrated approach promises to advance our capabilities in identifying breast cancer at its nascent stages, potentially improving patient outcomes significantly. In essence, our investigation reveals the transformative impact of machine learning on breast cancer diagnostics, urging continued exploration and adoption in clinical settings. It paves the way for a future where technology and healthcare converge to offer more precise, timely, and effective patient care.




REFERENCES

- [1] H. Heena *et al.*, "Knowledge, attitudes, and practices related to breast cancer screening among female health care professionals: a cross sectional study," *BMC Women's Health*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12905-019-0819-x.
- [2] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, Dec. 2016, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818560.
- [3] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Dec. 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.
- [4] F. G. -Gutierrez *et al.*, "Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms," *International Journal of Geriatric Psychiatry*, vol. 37, no. 2, Feb. 2022, doi: 10.1002/gps.5667.
- [5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [6] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2017, pp. 65–74, doi: 10.1145/3097983.3097997.
- [7] R. Kang, B. Park, Q. Ouyang, and N. Ren, "Rapid identification of foodborne bacteria with hyperspectral microscopic imaging and artificial intelligence classification algorithms," *Food Control*, vol. 130, Dec. 2021, doi: 10.1016/j.foodcont.2021.108379.
- [8] A. B. Yilmaz, Y. S. Taspinar, and M. Koklu, "Classification of malicious android applications using naive Bayes and support vector machine algorithms," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 2, pp. 269–274, 2022.
- [9] M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian," *Measurement*, vol. 72, pp. 32–36, Aug. 2015, doi: 10.1016/j.measurement.2015.04.028.
- [10] M. Mangukiyi, "Breast cancer detection with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 2, pp. 141–145, Feb. 2022, doi: 10.22214/ijraset.2022.40204.
- [11] A. R. Vaka, B. Soni, and S. R. K., "Breast cancer detection by leveraging machine learning," *ICT Express*, vol. 6, no. 4, pp. 320–324, Dec. 2020, doi: 10.1016/j.icte.2020.04.009.
- [12] V. N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer, 2000, doi: 10.1007/978-1-4757-3264-1.
- [13] W. Ali, "Hybrid intelligent android malware detection using evolving support vector machine based on genetic algorithm and particle swarm optimization," *International Journal of Computer Science and Network Security*, vol. 19, no. 9, pp. 15–28, 2019.
- [14] N. Gayathri and J. T. Selvi, "Investigation of various supervised machine learning algorithm to characterize mammogram breast images," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 2, pp. 406–417, 2023, doi: 10.17762/sfs.v10i2S.327.
- [15] P. Kaur, G. Singh, and P. Kaur, "Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification," *Informatics in Medicine Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.01.001.
- [16] H. N. Iqbal, A. B. Nassif, and I. Shahin, "Classifications of breast cancer diagnosis using machine learning," *International Journal of Computers*, vol. 14, pp. 86–86, Dec. 2020, doi: 10.46300/9108.2020.14.13.
- [17] Z. Kang, C. Catal, and B. Tekinerdogan, "Product failure detection for production lines using a data-driven model," *Expert Systems with Applications*, vol. 202, Sep. 2022, doi: 10.1016/j.eswa.2022.117398.
- [18] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravathy, "Prediction of brain stroke severity using machine learning," *Revue d'Intelligence Artificielle*, vol. 34, no. 6, pp. 753–761, Dec. 2020, doi: 10.18280/ria.340609.
- [19] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 18, no. 2, pp. 815–821, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14785.
- [20] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing*, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105941.
- [21] P. Misra and A. S. Yadav, "Impact of preprocessing methods on healthcare predictions," in *2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019, pp. 144–150, doi: 10.2139/ssrn.3349586.
- [22] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, 2020, doi: 10.1016/j.imu.2020.100402.
- [23] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," *Journal of Medical Systems*, vol. 43, no. 8, Aug. 2019, doi: 10.1007/s10916-019-1397-z.
- [24] Z. Peng, H. Wang, W. Wang, and Y. Jiang, "Intermodal transportation of full and empty containers in harbor-inland regions based on revenue management," *European Transport Research Review*, vol. 11, no. 1, Dec. 2019, doi: 10.1186/s12544-018-0342-4.
- [25] R. B. Asha and K. K. R. Suresh, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/j.gltp.2021.01.006.
- [26] A. Chawla, "Phishing website analysis and detection using machine learning," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 1, pp. 10–16, Mar. 2022, doi: 10.18201/ijisae.2022.262.
- [27] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.




BIOGRAPHIES OF AUTHORS

Soufyane Ayanouz    is currently a Ph.D. student at the Faculty of Sciences and Techniques of Tangier. He is also an engineer at the field of computer science since 2018. He received the engineering degree from Abdelmalek Essaadi University and he is the co-author of several papers published in IEEE Explorer, ACM, and in high indexed journals and conferences. He published several book chapters on Springer series, and is part of the reviewer's community. His key research relates to artificial intelligence and deep learning applied to the medical field. He can be contacted at email: ayanouz.soufiane@gmail.com.



Boudhir Anouar Abdelhakim    is currently an associate professor at the Faculty of Sciences and Techniques of Tangier. Actually, he was the president of the Mediterranean Association of Sciences and Technologies. He is an adviser at the Moroccan union against dropping out of school and IEEE Member since 2009. He received the HDR degree from Abdelmalek Essaadi University and he is the co-author of several papers published in IEEE Explorer, ACM, and in high indexed journals and conferences. He co-edited a several books published on Springer series and he is a co-founder of a series of international conferences. He can be contacted at email: boudhir.anouar@gmail.com.



Mohamed Ben Ahmed    is a full professor at the University Abdelmalek Essaadi. He received the Ph.D. degree in Computer Sciences and Telecommunications in 2010. He is currently a full Professor in computer sciences department at the Faculty of Sciences and Techniques - Morocco. He supervised several theses and conducted several international research projects about smart cities. He has authored more than 30 papers published in international journals and conferences. He co-edited a several books published on Springer series and he is a co-founder of a series of international conferences. He can be contacted at email: m.benahmed@gmail.com.